Recentered importance sampling with applications to Bayesian model validation

Ross Mcvinish, Kerrie Mengersen, Darfiana Nur, Judith Rousseau, Chantal Guihenneuc-Jouyaux

Recentered importance sampling with applications to Bayesian model validation

Ross McVinish[1], Kerrie Mengersen[2], Darfiana Nur[3], Judith Rousseau[4] and Chantal Guihenneuc-Jouyaux[5]

[1] University of Queensland , Brisbane, Queensland, Australia, 4001.

[2]QUT, Brisbane, Queensland, Australia, 4072.

[3] University of Newcastle, Callaghan, New South Wales, Australia, 2308.

[4] ENSAE - CREST and Université Paris Dauphine, Place du Maréchal De Lattre de Tassigny, Paris, France, 75016 and ENSAE-CREST (Paris).

[5] Université Paris Descartes, Faculté des sciences Pharmaceutiques et biologiques, Paris, France, 75006.

Abstract

Since its introduction in the early 90's, the idea of using importance sampling (IS) with Markov chain Monte Carlo (MCMC) has found many applications. This paper examines problems associated with its application to repeated evaluation of related posterior distributions with a particular focus on Bayesian model validation. We demonstrate that, in certain applications, the curse of dimensionality can be reduced by a simple modification of IS. In addition to providing new theoretical insight into the behaviour of the IS approximation in a wide class of models, our result facilitates the implementation of computationally intensive Bayesian model checks. We illustrate the simplicity, computational savings and potential inferential advantages of the proposed approach through two substantive case studies, notably computation of Bayesian p-values for linear regression models and simulation-based model checking. Supplementary materials including appendices and the R code for Section 3.1.2 are available online.

Keywords: curse of dimensionality, goodness of fit, importance sampling, p-values

1 Introduction

Bayesian model assessment typically requires the evaluation of multiple posterior distributions. As the posterior distribution for all but the simplest of models needs to be evaluated using Markov chain Monte Carlo (MCMC) or another computationally intensive technique, a complete model assessment can be computationally prohibitive. However, if the ratio of two posteriors can be computed quickly relative to an iteration of the MCMC algorithm, then one possibility to speed up computations is to use importance sampling (IS) with MCMC, as only a single MCMC run is needed to evaluate multiple posterior distributions. This idea has been used by a number of researchers for prior sensitivity analysis (Besag *et al.*, 1995; Nur *et al.*, 2009) and for leave-one-out cross-validation and case deletion diagnostics (Peruggia , 1997; Gerlach *et al.*, 1999; Stern and Cressie, 2000; Epifani *et al.*, 2008). We note that Peruggia (1997) and Epifani *et al.* (2008) have investigated quite precisely IS estimators for case deletion in some generalized linear models. They provide conditions to obtain finite moments for the IS weights, which is important for understanding the behavior of the IS approximation.

Despite these and numerous other successful applications of IS with MCMC, hereafter called MCMC-IS, there remain some issues which need further investigation. One case of particular interest is where it is necessary to evaluate multiple posterior distributions for which the prior is fixed but the datasets forming the posterior distributions are related but substantially different. This situation arises in evaluating certain Bayesian p-values such as those discussed in Dey *et al.* (1998); Bayarri and Berger (2000); Robins *et al.* (2000); Bayarri and Castellanos (2007); Fraser and Rousseau (2008) and in simulation studies evaluating the performance of a Bayesian estimator Gadja *et al.* (2010). In this situation, it is not clear that the original posterior distribution will provide an adequate

importance function.

In this paper, we propose using MCMC-IS when the posterior distribution is altered by a substantial change to the dataset. In Section 2, we study the asymptotic behavior of the IS weights, where the asymptotic is in terms of the sample size of the data. Although such an asymptotic analysis does not provide as precise a picture as in Peruggia (1997) or in Epifani *et al.* (2008), it has the advantage of being applicable to a wider class of models. Moreover, our asymptotic analysis highlights the detrimental influence of the dimension of the parameter space on the variance of the IS weights, a phenomenon generally referred to as the curse of dimensionality. This analysis provides theoretical support for using a simple location transformation (or link function) which successfully stabilises the importance weights even in moderate dimensions. Although such transformations were previously proposed by MacEachern and Peruggia (2000), their examples concerned only small changes to the posterior (case deletion in linear and nonlinear regression models). It was not clear that the approach would work for large changes. A further novelty of our work is the use of iterated IS to determine the location transformation. This can be particularly advantageous for models, such as generalized linear mixed effects models, where maximum likelihood estimation requires specialized techniques. Our results are then applied in Section 3 to two Bayesian model validation problems: (i) computation of Bayesian p-values for linear regression models, (ii) simulation-based model checking for a hierarchical logistic regression model. The paper concludes with some discussion of the advantages and disadvantages of the proposed methodology as well as possible future applications.

2 General approach and properties of the algorithm

2.1 Importance sampling approximation

Let $p(\theta|\mathbf{y})$ denote the posterior distribution for some sampling model $p(\mathbf{y}|\theta)$ and prior density $p(\theta)$ over parameter set Θ. Let $p(\theta|\mathbf{y}') \propto p(\mathbf{y}'|\theta)p(\theta)$ denote a second posterior density on Θ, where $\mathbf{y}'$ is a new dataset whose distribution is related to $\mathbf{y}$. For any integrable function h, let $I(h, \mathbf{y}) = \int_\Theta h(\theta)p(\theta|\mathbf{y})d\theta$. From the usual IS approach $I(h, \mathbf{y}') = I(hp(\cdot|\mathbf{y}')/p(\cdot|\mathbf{y}), \mathbf{y})$ if the support of $p(\theta|\mathbf{y}')$ is included in the support of $p(\theta|\mathbf{y})$, which we assume hereafter. Let $(\theta_t)_{t=1}^T$ be a Markov chain with stationary distribution $p(\theta|\mathbf{y})$. Typically, $p(\theta|\mathbf{y})$ is only known up to a normalizing constant so that the IS approximation, based on the Markov chain $(\theta_t)_{t=1}^T$, is

$$\hat{I}_T(h, \mathbf{y}') \quad = \quad \frac{\sum_{t=1}^T h(\theta_t)w(\theta_t)}{\sum_{t=1}^T w(\theta_t)}, \quad \text{where} \quad w(\theta) = \frac{p(\mathbf{y}'|\theta)}{p(\mathbf{y}|\theta)}. \tag{1}$$

It is well known that if (θ_t) is ergodic and if $\int_\Theta |h(\theta)|p(\theta|\mathbf{y}')d\theta < +\infty$, then $\hat{I}_T(h, \mathbf{y}')$ converges almost surely to $I(h, \mathbf{y}')$ as $T \to \infty$ (Smith and Roberts , 1993). More precise results on the accuracy of the resulting estimates depend heavily on the behavior of the Markov chain and we refer to Nur *et al.* (2009) for a discussion on conditions implying a central limit theorem on $\sqrt{T}\left(\hat{I}_T(h, \mathbf{y}') - I(h, \mathbf{y}')\right)$.

An important issue in IS concerns the variability of the weights $w(\theta)$ since it determines the rate of convergence of $\hat{I}_T(h, \mathbf{y}')$. It is well known that the variability of the weights increases exponentially with the dimension of Θ. This fact is made more precise in the following subsection.

2.2 Variability of the weights

For most models, both $p(\theta|\mathbf{y})$ and $p(\theta|\mathbf{y}')$ converge to a dirac mass at some $\theta_0 \in \Theta$ as the amount of data increases. Despite this, it is not obvious that $p(\theta|\mathbf{y})$ will be a useful importance function for $p(\theta|\mathbf{y}')$ even in the large data setting. This is because the two distributions may have very little overlap for any finite amount of data. To address this question, we examine the distribution of the normalized importance weights

$$\tilde{w}(\theta) = \frac{w(\theta)}{E(w(\theta|\mathbf{y}))} = \frac{p(\theta|\mathbf{y}')}{p(\theta|\mathbf{y})}, \quad \text{where } E(w(\theta)|\mathbf{y}) = \int_\Theta w(\theta) dp(\theta|\mathbf{y}),$$

as the amount of data increases. The necessary modifications to extend the following results to self-normalized weights are given in the appendix.

We consider situations in which the posterior distribution $p(\theta|\mathbf{y})$ and $p(\theta|\mathbf{y}')$ can be approximated by Gaussian distributions centered at $\hat{\theta}^{\mathbf{y}}, \hat{\theta}^{\mathbf{y}'}$ and with asymptotic covariance matrices $J(\mathbf{y})^{-1}$ and $J(\mathbf{y}')^{-1}$. Although not strictly necessary, we assume that the size of both datasets $\mathbf{y}$ and $\mathbf{y}'$ is n. This assumption is satisfied for the examples in Section 3 such as the computation of Bayesian p-values as defined in Bayarri and Berger (2000) and Fraser and Rousseau (2008). Assume also that $J(\mathbf{y}) = nI_0(1 + o_P(1))$ and $J(\mathbf{y}') = nI_0(1 + o_P(1))$, where I_0 is a fixed positive definite matrix. Situations where the Gaussian approximation hold are quite common. For example, the Gaussian approximation holds for finite dimensional settings ($\Theta \subset \mathbb{R}^r$, $r \geq 1$) where the data are independent realisations from a regular model (Kass *et al.*, 1989) with $\mathbf{y}$ and $\mathbf{y}'$ having very similar distributions. However, independence and regularity are not necessary conditions, see for instance Philippe and Rousseau (2003) for dependent data and Ghosal and Samanta (1997) for non regular models. Using the Gaussian approximation of the posterior we

have,

$$\begin{aligned}
\tilde{w}(\theta) &= \exp\left\{-\frac{(\hat{\theta}^{\mathbf{y}'}-\hat{\theta}^{\mathbf{y}})^t J(\mathbf{y}')(\hat{\theta}^{\mathbf{y}'}-\hat{\theta}^{\mathbf{y}})}{2}\right\}\exp\left\{-(\theta-\hat{\theta}^{\mathbf{y}})^t J(\mathbf{y}')(\hat{\theta}^{\mathbf{y}}-\hat{\theta}^{\mathbf{y}'})\right\} \\
&\quad \times \exp\left\{-\frac{(\theta-\hat{\theta}^{\mathbf{y}})^t\left(J(\mathbf{y}')-J(\mathbf{y})\right)(\theta-\hat{\theta}^{\mathbf{y}})}{2}\right\}(1+o_P(1)) \qquad (2) \\
&= \exp\left\{-\frac{n(\hat{\theta}^{\mathbf{y}'}-\hat{\theta}^{\mathbf{y}})^t I_0(\hat{\theta}^{\mathbf{y}'}-\hat{\theta}^{\mathbf{y}})}{2} - n(\theta-\hat{\theta}^{\mathbf{y}})^t I_0(\hat{\theta}^{\mathbf{y}}-\hat{\theta}^{\mathbf{y}'})\right\}(1+o_P(1)) \qquad (3)
\end{aligned}$$

Note that (3) holds even when $\mathbf{y}$ and $\mathbf{y}'$ are not independent since it was derived from marginal Laplace approximations of $p(\theta|\mathbf{y})$ and $p(\theta|\mathbf{y}')$. Note also that the results remain valid if we only assume that $\mathbf{y}'$ has dimension n' with $n'/n = 1 + o(1)$.

A useful summary of the distribution of the normalized weights is its variance. Under additional conditions on the integrability of $w(\theta)^2$ which are given in the Appendix, we may express the asymptotic variance as

$$\mathrm{var}\left(\tilde{w}(\theta) \mid \mathbf{y}, \mathbf{y}'\right) = \exp\left\{n(\hat{\theta}^{\mathbf{y}'}-\hat{\theta}^{\mathbf{y}})^t I_0(\hat{\theta}^{\mathbf{y}'}-\hat{\theta}^{\mathbf{y}}) + o_P(1)\right\} - 1. \qquad (4)$$

The stability of the weights can thus be predicted using the statistic

$$\Delta(\mathbf{y}, \mathbf{y}') = (\hat{\theta}^{\mathbf{y}'}-\hat{\theta}^{\mathbf{y}})^t J(\mathbf{y}')(\hat{\theta}^{\mathbf{y}'}-\hat{\theta}^{\mathbf{y}'}).$$

The centers $\hat{\theta}^{\mathbf{y}}$ and $\hat{\theta}^{\mathbf{y}'}$ can be the posterior means, maximum likelihood estimates or maximum a posteriori estimates, depending on the ease of computation. To understand better the possible impact of the dimension r of Θ, assume that each sample $\mathbf{y}$ and $\mathbf{y}'$ comprises independently and identically distributed observations from a regular model. If $\mathbf{y}$ and $\mathbf{y}'$ are mutually independent, then asymptotic normality of the maximum likelihood estimator implies that $\Delta(\mathbf{y}, \mathbf{y}') \xrightarrow{d} 2\chi^2_r$, where r is the dimension of θ. This implies that the variance of the weights is asymptotically (in n) exponentially increasing in the dimension r of the parameter space. This situation is the least favorable as far as the variability of the weights is concerned. However, in the application to Bayesian p-values

as defined in Section 3, $\mathbf{y}$ and $\mathbf{y}'$ are not independent. In that case, conditional on $(\hat{\theta}_l^{\mathbf{y}'} = \hat{\theta}_l^{\mathbf{y}}, l = 1, \ldots, s < r)$, $\mathbf{y}$ and $\mathbf{y}'$ are two independent and identically distributed vectors. This leads to $\Delta(\mathbf{y}, \mathbf{y}') \xrightarrow{d} 2\chi^2_{r-s}$ (Gouriéroux and Monfort, 1996) and the variance of the weights remains exponentially increasing with $r - s$.

Relation (3) suggests that a simple transformation may significantly stabilize the weights, at least when the size of the data is large compared to the dimension of the parameter. The idea of transforming the values $(\theta_t)_{t=1}^T$ before computing the importance weights using an importance link function has been proposed in MacEachern and Peruggia (2000), but the above calculations imply that a very simple transformation is often quite effective. For all t set

$$\theta'_t = \theta_t + \hat{\theta}^{\mathbf{y}'} - \hat{\theta}^{\mathbf{y}}, \tag{5}$$

with modified weights given by

$$w'(\theta_t) = \frac{p(\theta'_t)p(\mathbf{y}'|\theta'_t)}{p(\theta_t)p(\mathbf{y}|\theta_t)}, \quad \tilde{w}'(\theta_t) = \frac{w'(\theta_t)}{E[w'(\theta)|\mathbf{y}]}.$$

Again applying the Laplace approximation as in equation (3), we see $\tilde{w}'(\theta) \xrightarrow{p} 1$. Furthermore, under the conditions given in the Appendix, we obtain that $\mathrm{var}(\tilde{w}'(\theta)|\mathbf{y}') = o_P(1)$. In equation (5) we have assumed that $\Theta = \mathbb{R}^r$. If $\Theta \subset \mathbb{R}^r$, then the recentering will need to be applied after Θ has been mapped to all of $\mathbb{R}^r$. For example, if $\Theta = [0, \infty)$ then the recentering (5) is applied after taking a log transformation.

It is of interest to note that the naive MCMC-IS approach suffers from the curse of dimensionality even when the size of the data is much larger than the dimension and that the simple linear transform (5) stabilizes the weights, whatever the dimension, provided it is much smaller than the number of observations.

To apply transformation (5), one needs to be able to compute the centering points $\hat{\theta}^{\mathbf{y}'}$ and $\hat{\theta}^{\mathbf{y}}$. Often the maximum likelihood estimate and the maximum a posteriori estimate

are not available in tractable analytic forms and so the centering procedure becomes difficult to implement. Furthermore, the posterior mean will be a better center point for some posterior distributions such as a Gaussian posterior distribution. We thus propose an iterative algorithm to compute the posterior mean centered MCMC-IS.

- (1) Sample $(\theta_t)_{t=1}^T$ from the posterior distribution $p(\theta|\mathbf{y})$ using MCMC and compute

$$w(\theta_t), \quad \tilde{\theta}_T^{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\theta_t, \quad \text{and} \quad \tilde{\theta}_T^{\mathbf{y}'} = \frac{\sum_{t=1}^T \theta_t w(\theta_t)}{\sum_{t=1}^T w(\theta_t)}$$

- (2) Set $\theta_t' = \theta_t + \tilde{\theta}_T^{\mathbf{y}'} - \tilde{\theta}_T^{\mathbf{y}}$ and compute the modified importance weights $w'(\theta_t)$.

- (3) Update the estimate of the posterior mean by

$$\tilde{\theta}_T^{\mathbf{y}'} = \frac{\sum_{t=1}^T \theta_t' w'(\theta_t)}{\sum_{t=1}^T w'(\theta_t)}$$

  and return to step (2) if the estimate of the posterior mean has not yet stabilized.

Based on the simulations in section 3, two iterations of the algorithm are often sufficient in order to stabilize the posterior mean estimate.

We could also consider more complicated importance link functions, such as the location-scale transformation of θ_t suggested by (2),

$$\theta_t' = J(\mathbf{y}')^{1/2} J(\mathbf{y})^{-1/2}(\theta_t - \hat{\theta}^{\mathbf{y}}) + \hat{\theta}^{\mathbf{y}'}.$$

However, as shown with equation (3), the scale factor does not play as crucial a role as the location transformation, at least when the number of observations is much larger than the dimension. Note also that, although the re-centering is motivated by asymptotic normality of the posterior, it is a more robust approach than a simple IS approximation based on a Gaussian proposal, which would require strong tail conditions to be efficient. In our framework, the re-centering can be quite efficient even though the tails of the target distribution are much heavier than Gaussian tails.

3 Applications

In this section we apply the MCMC-IS algorithms described in the previous section to three different Bayesian model validation settings; the computation of Bayesian p-values for hypothesis testing on parameters, a simulation-based approach to checking proper priors in a hierarchical mode. In the first subsection, Bayesian p-values are used to test the significance of certain coefficients in a Gaussian linear regression model. Examples are given using both a standard noninformative prior and the Bayesian lasso prior. The second subsection concerns detecting incompatibility between the hierarchical priors of a logistic regression model and the data using simulation-based model checking.

Although the models studied in this section are relatively simple, they are often used in applications. Furthermore, despite the simplicity of these models, the application of certain Bayesian model validation techniques, such as those considered here, remains computationally challenging due to the need to evaluate hundreds or even thousands of posterior distributions. The results obtained here suggest that recentered MCMC-IS could be used for Bayesian model validation of models for which performing hundreds of MCMC runs is not possible.

3.1 Parametric regression and computations of Bayesian p-values

Consider the normal linear model

$$\mathbf{y} \sim \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_{ip} + \epsilon,$$

where the $\epsilon \sim N(0, \sigma^2 I_n)$ and I_n is the $(n \times n)$ identity matrix. Let $\theta = (\beta_0, \ldots, \beta_p, \sigma^2)$ so that $r = p+2$. We are interested in the problem of determining if the subset of covariates $(\mathbf{x}_s, \ldots, \mathbf{x}_p)$, with $s \leq p$ has any effect on the response variable $\mathbf{y}$. This is equivalent to

the hypothesis test

$$H_0 : \beta_s = \cdots = \beta_p = 0 \quad \text{against} \quad H_1 : \beta_i \neq 0 \text{ for at least one } i \in \{s, \ldots, p\}. \tag{6}$$

Let $\psi = (\beta_0, \ldots, \beta_{s-1}, \sigma^2) \in \Psi$ be the parameter under H_0 and let $\mathbf{y}_o$ be the observed data. One approach to this problem is to consider a p-value defined by $\Pr(t(\mathbf{y}) > t(\mathbf{y}_o))$, where $t(\mathbf{y}_o)$ is the (Bayesian) test statistic. The choice of probability measure under H_0 leads to different types of p-values such as prior/posterior predictive p-values and plug-in p-values (Bayarri and Berger, 2000; Robins *et al.*, 2000). Here we follow Robert and Rousseau (2002) and Fraser and Rousseau (2008) and take the probability measure to be

$$m(\mathbf{y} \mid \hat{\psi}^{\mathbf{y}_o}) = \int p(\mathbf{y} \mid \hat{\psi}^{\mathbf{y}} = \hat{\psi}^{\mathbf{y}_o}, \psi) p(\psi \mid \hat{\psi}^{\mathbf{y}_o}) d\psi, \tag{7}$$

where $p(\psi \mid \hat{\psi}^{\mathbf{y}_o})$ is the distribution of ψ conditional on the maximum likelihood estimate (MLE) $\hat{\psi}^{\mathbf{y}_o}$ under H_0 and $p(\mathbf{y} \mid \hat{\psi}^{\mathbf{y}} = \hat{\psi}^{\mathbf{y}_o}, \psi)$ is the distribution under H_0 of a new sample $\mathbf{y}$ conditional on the MLE from this sample being equal to the MLE from the observed data. This leads to the conditional predictive p-value. In the Gaussian case, the conditional distribution of $\mathbf{y}$ given $\hat{\psi}^{\mathbf{y}}$ does not depend on ψ. One can then show that $m(\mathbf{y} \mid \hat{\psi}^{\mathbf{y}_o})$ has the uniform distribution on an ellipsoid.

3.1.1 Standard non-informative prior

For illustration we set $r = 1$ in the hypotheses (6) and take the test statistic to be the posterior expectation of $\sum_{i=1}^{p} \beta_i^2$. Adopting the standard non-informative prior $\pi(\theta) \propto 1/\sigma^2$, the test statistic can be expressed as

$$t(\mathbf{y}) = \mathbb{E}\left(\sum_{i=1}^{p} \beta_i^2 \mid \mathbf{y}, X \right) = \sum_{i=1}^{p} \hat{\beta}_i(\mathbf{y})^2 + \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p - 2} tr\left[(X^t X)^{-1} \right], \tag{8}$$

where $\hat{\beta}_i(\mathbf{y})$ is the ordinary least squares estimate of β_i and $\hat{y}_i$ is the i-th element of $X\hat{\beta}(\mathbf{y})$ and X is the matrix of covariates whose i-th row corresponds to the i-th individual. Thus,

an accurate approximation to the p-value can be obtained from $M^{-1}\sum_{i=1}^{M}\mathbb{1}(t(\mathbf{y}_i) > t(\mathbf{y}_o))$, where $\mathbf{y}_i$ are independent realisations from (7) and the test statistic is evaluated from (8). We will refer to p-values obtained in this way as 'exact'.

In the simulations that follow we compare the accuracy of p-values obtained using five different approximations to the test statistic (8). The approximations are:

[1] MCMC: For each realization $\mathbf{y}_i$ from (7), the test statistic is evaluated by first generating a sample from the resulting posterior distribution using Gibbs sampling and then computing the average of $\sum_{i=1}^{p}\beta_i^2$.

[2] MCMC-IS: MCMC is used to generate a sample from the posterior distribution $p(\theta|\mathbf{y}_1)$ and compute $t(\mathbf{y}_1)$. For $i \geq 2$, $t(\mathbf{y}_i)$ is computed by MCMC-IS as described in Section 2.

[3] MLE centered MCMC-IS: MCMC is used to generate a sample from the posterior distribution $p(\theta|\mathbf{y}_1)$ and compute $t(\mathbf{y}_1)$. For $i \geq 2$, $t(\mathbf{y}_i)$ is computed by MCMC-IS with the centering transformation

$$\beta_t' \;=\; \beta_t + \hat{\beta}^{\mathbf{y}_i} - \hat{\beta}^{\mathbf{y}_1}, \quad \log(\sigma_t') = \log(\sigma_t) + \log(\hat{\sigma}^{\mathbf{y}_i}) - \log(\hat{\sigma}^{\mathbf{y}_1}).$$

[4] Posterior mean centered MCMC-IS: Similar to [3], except the MLE is replaced by the posterior mean which is estimated by IS.

[5] Iterated posterior mean centered MCMC-IS: Similar to [4], except that the posterior mean is obtained by 2 iterations of the algorithm given in Section 2.2.

To illustrate the effect of dimension and how the recentering diminishes this effect, we consider three cases: the first has one covariate under H_1, the second has 9 covariates under H_1 and the third has 24 covariates under H_1. In the simulation study, samples

of $n = 250$ observations are generated under $H_0 : \ (\beta_0 = 1, \sigma^2 = 1)$ 250 times. For each p-value, $M = 1000$ samples are generated from (7) on which the test statistics are computed. We simulate from the posterior distributions of θ using a standard Gibbs sampling algorithm which is run for 10^5 iterations with a 1% burn-in. The results are displayed in Figures 1– 3.

Figures 1–3 show that MLE centered MCMC-IS performs almost as well as MCMC in approximating the p-values. It also shows that simple MCMC-IS displays considerable variability in small dimensions and in moderate dimensions the results are too biased to be of use. It is intersting that posterior mean centered MCMC-IS performs relatively well with $p = 1$, but its performance deteriorates appreciably as p increases to 9 and with $p = 24$ the results are very biased. Iterated posterior mean centered MCMC-IS displays the same type of behaviour however the rate at which its performance deteriorates is much slower. It is probable that the performance of iterated posterior mean centered MCMC-IS could be improved by further iteration. One suprising result from the simulations is that the performance of MCMC and MLE centered MCMC-IS appears to improve as the dimension increases from $p = 1$ to $p = 9$ and no deterioration is observed for $p = 24$. Our explanation for this is that in these cases the test statistic is small relative to the sampling error in approximating the test statistic. When the dimension increases, the test statistic typically increases which results in an improvement of the relative error.

3.1.2 The Bayesian lasso

The Bayesian lasso is a linear regression model where the coefficients $\beta_1, \ldots, \beta_p$ are given (conditionally) independent, mean zero, Laplace prior distributions. This results in considerably greater shrinkage of the regression coefficients compared with the standard

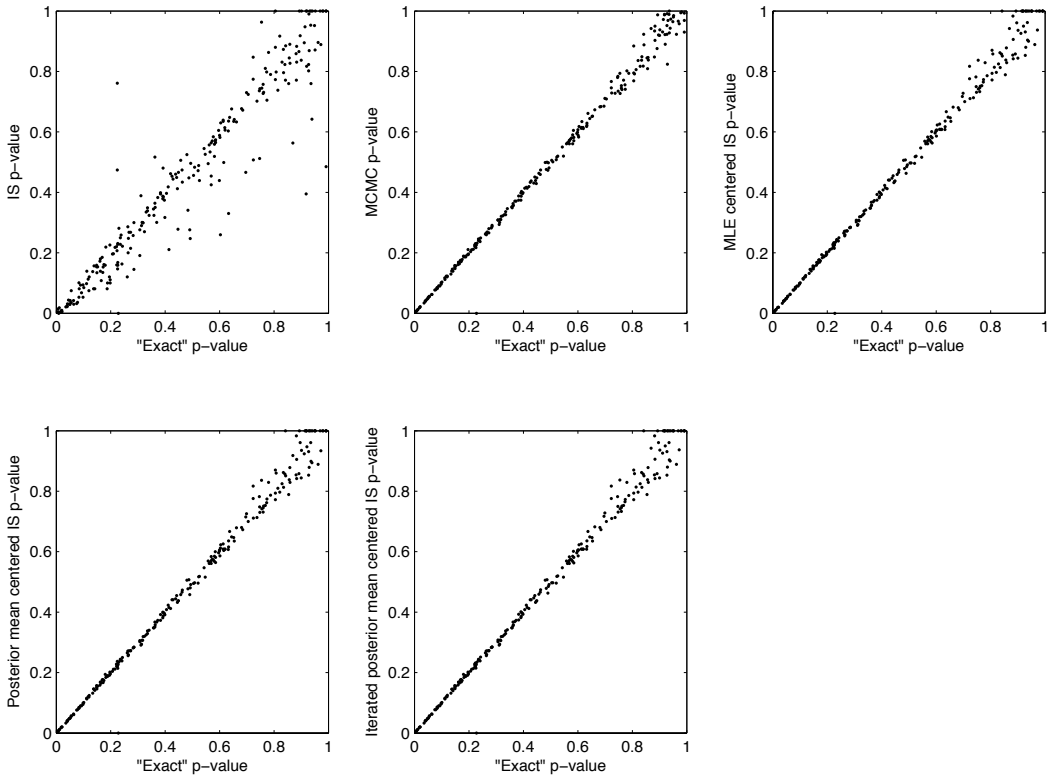

Figure 1: Scatter plots of 'Exact' p-values against approximated p-values where H_1 includes one covariate.

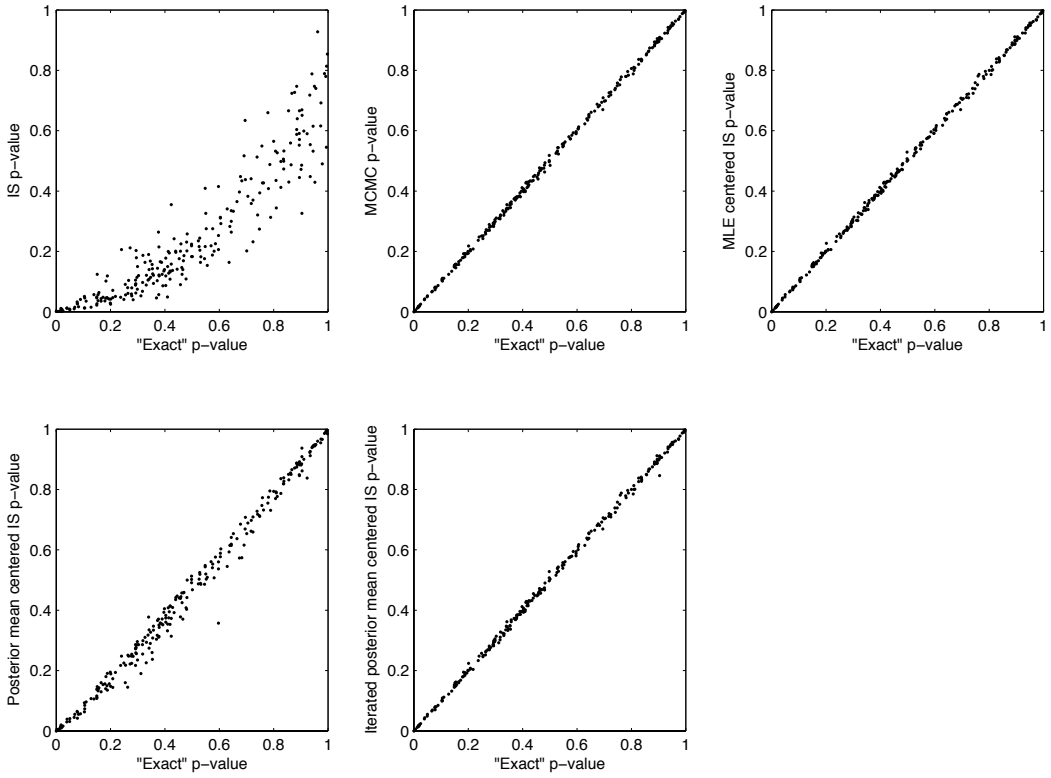

Figure 2: Scatter plots of 'Exact' p-values against approximated p-values where H_1 includes nine covariates.

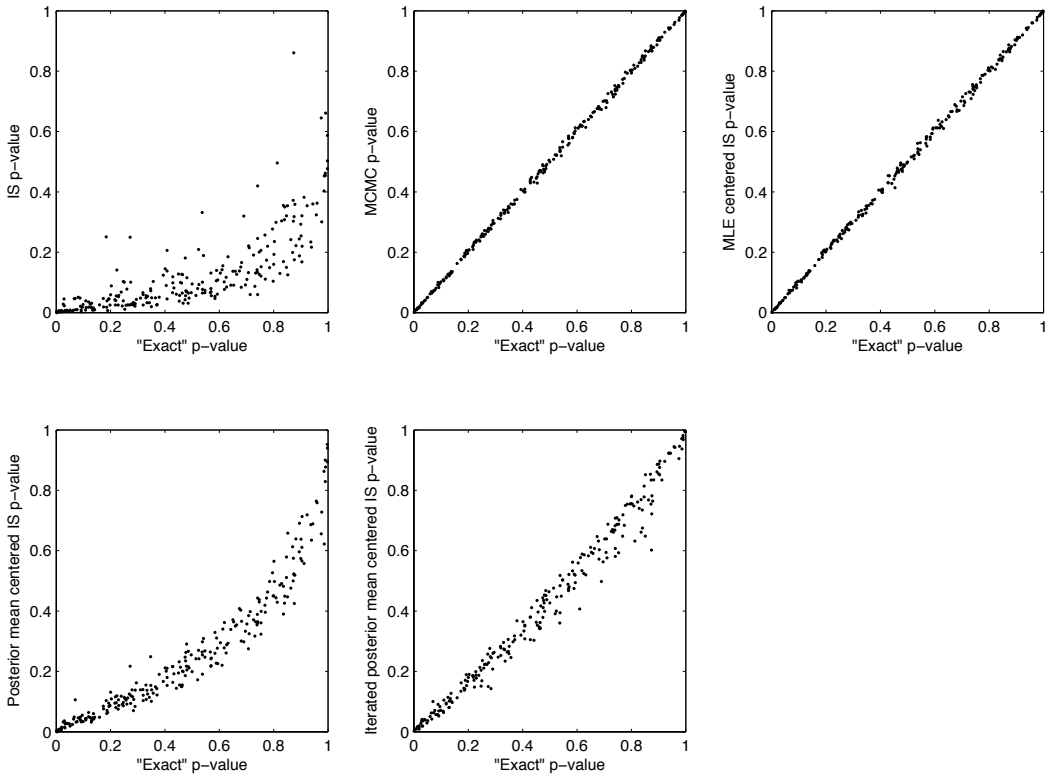

Figure 3: Scatter plots of 'Exact' p-values against approximated p-values where H_1 includes twenty four covariates.

Gaussian prior. Following Park and Casella (2008), we take the prior for the regression model to be

$$\begin{aligned} p(\beta_0, \ldots, \beta_p|\sigma^2\lambda^2) &= \prod_{j=1}^{p} \frac{\lambda}{2\sigma} \exp\left(-\lambda|\beta_j|/\sigma\right) \\ p(\sigma^2, \lambda^2) &= \frac{1}{\sigma^2} \times \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} \exp\left(-\delta\lambda^2\right), \quad (r = 1, \delta = 1.78). \end{aligned}$$

Note that β_0 has the improper, uniform prior on $\mathbb{R}$. Park and Casella (2008) proposed a Gibbs sampler based on the scale mixture of normals representation of the Laplace distribution. An alternative Gibbs sampler is presented in Hans (2009). In their analysis of the Diabetes dataset (Efron *et al.*, 2004) where $p = 10$ and $n = 442$, Park and Casella (2008) showed that the 95% credible intervals for a number of the regression coefficients contain zero (variables age, tc,ldl, hdl, tch and glu). We illustrate the computation of the Bayesian p-value for the hypothesis test that regression coefficients of these five variables

are simultaneously zero.

As the posterior distribution of the Bayesian lasso is not Gaussian, we cannot compute the test statistics exactly. Instead we need to use MCMC, MCMC- IS or one of the variants of MCMC-IS considered in the previous section. MCMC-IS and its variants offer considerable computational savings, since sampling from the full conditional of $(\beta_1, \ldots, \beta_p)$ requires solution of a p-dimensional linear system (for the mean of the multivariate normal) and inverting a $p \times p$ matrix (for the covariance matrix of the multivariate normal). To approximate the p-values 1000 test statistics were generated. In the MCMC, a chain length of 10^5 was used with a burn-in of 1%.

The results displayed in Figure 4 indicate that centering is necessary and that iterated posterior mean centering gives superior results. The estimated p-values in this example were 0.048 (MCMC), 0.091 (MCMC-IS), 0.047 (Posterior mean centered MCMC-IS) and 0.045 (Iterated posterior mean centered MCMC-IS). Although the p-value from Posterior mean centered MCMC-IS is slightly more accurate in this instance, from Figure 4 we believe that this is due to random variation and that iterated posterior mean centered MCMC-IS should be preferred.

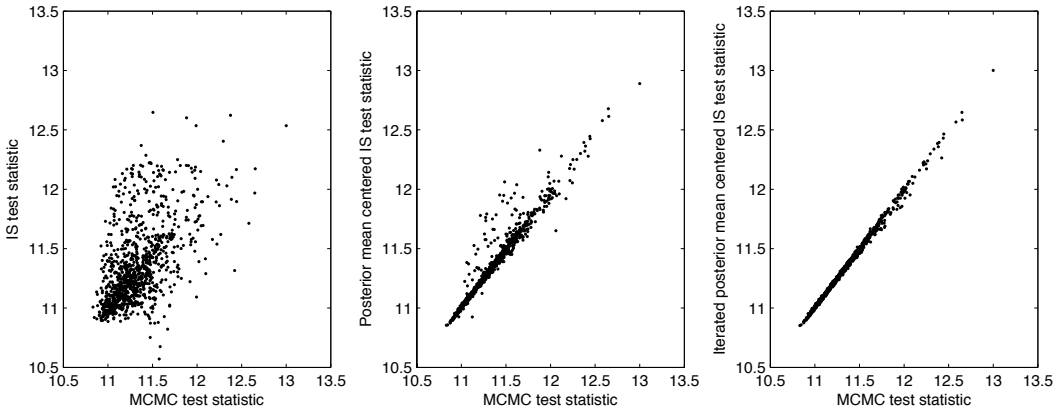

Figure 4: Scatter plots of MCMC test statistics against approximated test statistics computed using MCMC-IS and its variants. The test statistic for the data is approximately 11.94.

3.2 Simulation-based model checking

Dey *et al.* (1998) proposed a computationally intensive method for checking the suitability of proper priors used in a hierarchical model. Bayarri and Castellanos (2007) concluded that the approach works well in detecting incompatibility between the model and data when proper priors are used. A summary of the steps involved in the method is given below:

1. For $r = 1, \ldots, R$ repeat

   (a) Simulate $\mathbf{y}^{(r)}$ from the prior predictive distribution, that is, simulate $\theta^{(r)} \sim p(\theta)$ then simulate $\mathbf{y}^{(r)} \sim p(\mathbf{y}|\theta^{(r)})$.

   (b) For discrepancy measure of interest $d(\mathbf{y}, \theta)$, compute the vector of quantiles $\mathbf{q}^{(r)} = (q^{(r)}_{0.05}, q^{(r)}_{0.25}, q^{(r)}_{0.5}, q^{(r)}_{0.75}, q^{(r)}_{0.95})$ where $q^{(r)}_{\alpha}$ denotes the α quantile of the posterior distribution $p(d(\mathbf{y}^{(r)}, \theta)|\mathbf{y}^{(r)})$.

2. Compute the vector $\bar{\mathbf{q}}$ of averages over r and for the original data $\mathbf{y}^{(0)}$ compute $\mathbf{q}^{(0)}$.

3. Compute the $R + 1$ Euclidean distances between $\mathbf{q}^{(r)}$ and $\bar{q}$.

4. Perform a one-sided, upper tail Monte Carlo test comparing the distance between $\mathbf{q}^{(0)}$ and $\bar{\mathbf{q}}$ with the distances between $\mathbf{q}^{(r)}$ and $\bar{\mathbf{q}}$, $r = 1, \ldots, R$.

In this subsection, we study the performance of the MCMC-IS methods described in subsection 2.2 in estimating the quantiles of the posterior distributions $p(d(\mathbf{y}^{(r)}, \theta)|\mathbf{y}^{(r)})$ described in Step 1.

The MCMC-IS methods are applied to simulation-based model checking for the heart transplant dataset and model from section 5 of Dey *et al.* (1998). This dataset contains the number of patients developing problems leading to short term organ rejection and

the total number of transplant patients at 10 centers and grouped into 5 age groups. The number of patients developing problems at center i and in age group j, denoted y_{ij} is modelled as $y_{ij} \sim B(p_{ij}, n_{ij})$ where

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + \beta_i x_j,$$

with ages x_j have been centered and scaled. The prior on (α_i, β_i) is given by

$$\begin{aligned} \alpha_i &\sim N(\mu_\alpha, \tau_\alpha^2), \qquad \beta_i \sim (\mu_\beta, \tau_\beta^2), \\ \mu_\alpha &\sim N(-0.9, (0.2)^2), \qquad \mu_\beta \sim N(0.17, (0.05)^2), \\ \tau_\alpha^{-2} &\sim Gamma(2.16, 0.0464), \qquad \tau_\beta^{-2} \sim Gamma(2.006944, 0.002517361). \end{aligned}$$

We use two sets of discrepancy measure;

$$d_1^{ij} = y_{ij} - \frac{n_{ij}\exp(\alpha_i + \beta_i x_j)}{1+\exp(\alpha_i + \beta_i x_j)}, \quad \text{and} \quad d_{2|1}^{ij} = y_{ij} - \frac{n_{ij}\exp(\mu_\alpha + \mu_\beta x_j)}{1+\exp(\mu_\alpha + \mu_\beta x_j)}.$$

For this simulation study we took $R = 250$; the length of the MCMC chains was 10^4 and a 10% burn-in was used in each case. For each $r = 2, \ldots, R$, we estimate $\mathbf{q}^{(r)}$ for each of the d_1^{ij} and $d_{2|1}^{ij}$, $i = 1, \ldots, 10,\ j = 1, \ldots, 5$, using MCMC, MCMC-IS, posterior mean centered MCMC-IS and iterated posterior mean centered MCMC-IS with 2 and 5 iterations. The accuracy of the results from the MCMC-IS methods is quantified using the empirical correlation coefficients between these estimates and the estimates obtained from MCMC. This yields a sample of $R-1$ correlation coefficients for each of the MCMC-IS methods. Summaries of these samples are given in Table 3.2.

As in the previous examples, centering the samples using the posterior mean before applying IS results in a considerable improvement in accuracy compared to direct application of IS. Using iterated posterior mean centering MCMC-IS yields further improvement over posterior mean centered MCMC-IS. Although more accurate estimates are obtained

Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
MCMC-IS	0.8054	0.9760	0.9887	0.9820	0.9953	0.9997
Posterior mean centered MCMC-IS	0.8432	0.9933	0.9981	0.9937	0.9996	1.0000
Posterior mean centered MCMC-IS (2 iterations)	0.8843	0.9985	0.9996	0.9976	0.9999	1.0000
Posterior mean centered MCMC-IS (5 iterations)	0.9727	0.9993	0.9997	0.9994	0.9999	1.0000

Table 1: Summary statistics of correlation coefficients between the estimates obtained from MCMC-IS methods with estimates obtained by MCMC.

by further iteration of the procedure, the subsequent improvements are not as large. Also, the gains in accuracy need to be balanced against the amount of additional computation required.

4 Summary and Conclusion

Importance sampling has previously been shown to be useful in situations requiring the computation of expectations or decisions with respect to multiple posterior distributions. In this paper we have focused on the properties of IS when applied to Bayesian model validation techniques.

In this paper, we make two important contributions to the literature. Firstly, we provide an examination of the IS weights in a setting that is relevant to Bayesian model validation techniques. Secondly, we demonstrate that, in this setting, recentering can stabilize the IS weights for moderatly complex models. The required recentering can be

estimated easily and accurately using iterated IS. These contributions are demonstrated in two Bayesian model validation settings: calculation of Bayesian p-values for linear regression models and checking of priors for a hierarchical logistic regression model.

We note that the full gains that can be anticipated from the proposed approach are not realized in the examples used in this paper, since for comparative purposes we needed to select situations in which MCMC could also be used. However, the results are still compelling. The greatest computational gain will be achieved in situations where an iteration of the MCMC algorithms is computationally expensive relative to the evaluation of the likeihood and prior. In the Bayesian lasso example, the Gibbs sampling algorithm involved simulation of a p-dimensional multivariate normal random variable which requires $O(p^3)$ operations whereas the most expensive part of evaluating the likelihood is the solution of a p-dimensional linear system which requires $O(p^2)$ operations. Although these examples concerned model checking, the same ideas can be used when checking for the validity of a Bayesian methodology through a simulation study, as considered in Gadja *et al.* (2010).

Our hope is that the theory and applications presented here will convince the reader that computationally intensive model validation techniques can be made feasible using recentered IS.

5 Supplemental Materials.

The supplemental materials are contained in single archive, they are composed of the following two items

Appendix In this Appendix, we give the proof and conditions for (4) (Section 1.1) together with modifications to the arguments presented in Section 2.2 to apply to

self-normalized importance weights.

R code for the Lasso example In this supplementary material with provide the R code used to perform the Lasso example presented in Section 3.1.2.

References

Bayarri, M.J., and Berger, J.O. (2000) P values for composite null models, *Journal of the American Statistical Association*, 95, 1127-1142.

Bayarri, M.J., and Castellanos, M.E. (2007) Bayesian checking of the second levels of hierarchical models, *Statistical Science*, 22, 322-343.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion)," *Statistical Science*, 10, 3-66.

Dey, D., Gelfand, A.E., Swartz, T.B. and Vlachos, A.K. (1998) A simulation-intensive approach for checking hierarchical models. *Test*, 7, 325-346.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression, *Annals of Statistics*, 32, 407-499.

Epifani, I., MacEachern, S. and Peruggia, M. (2008) Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, 2, 774-806.

Fraser, D. and Rousseau, J. (2008) Studentization and deriving accurate p-values. *Biometrika*, 95, 1-16.

Gadja, D., Guihenneuc-Jouyaux, C., Rousseau, J., Mengersen, K. and Nur, D. (2010) Use in practice of importance sampling for repeated MCMC for Poisson models, *Electronic Journal of Statistics*, 4, 361-383.

Gerlach, R., Carter, C., and Kohn, R. (1999) Diagnostics for time series analysis, *Journal of Time Series Analysis*, 20, 309-330.

Ghosal, S. and Samanta, T. (1997) Asymptotic expansions of posterior distributions in nonregular cases *Annals of the Institute of Mathematical Statistics*, 49, 181-197.

Gouriéroux, C. and Monfort, A. (1996) *Statistique et Modèles Econométriques*, volume 1, Economica.

Hans, C. (2009) Bayesian lasso regression, *Biometrika*, 96, 835-845.

Kass, R., Tierney, L., and Kadane, J.B. (1989) Approximate methods for assessing influence and sensitivity in Bayesian analysis, *Biometrika*, 76, 663-74.

MacEachern, S.N. and Peruggia, M. (2000) Importance link function estimation for MCMC methods. *Journal of Computational and Graphical Statistics*, 9, 99-121.

Nur, D., Allingham, D., Rousseau, J., Mengersen, K. and McVinish R. (2009) Bayesian hidden Markov model for DNA sequence segmentation : A prior sensitivity analysis, *Computational Statistics and Data Analysis*, 53, 1873-1882.

Park, T. and Casella, G. (2008) The Bayesian lasso, *Journal of the American Statistical Society*, 103, 681-686.

Peruggia, M. (1997) On the variability of case-deletion importances sampling weights in the Bayesian linear model, *Journal of the American Statistical Association*, 92, 199-207.

Philippe, A. and Rousseau, J. (2003) Non-informative priors for Gaussian long-memory processes, *Bernoulli*, 8, 451-473.

Pollard, D. (1984) *Convergence of stochastic processes.* Springer, New York.

Robert, C.P., and Rousseau, J. (2002) A mixture approach to Bayesian goodness of fit, *Cahiers du CEREMADE* 2002-9.

Robins, J.M., van der Vaart, A., and Ventura, V. (2000) Asymptotic distribution of P values in composite null models, *Journal of the American Statistical Association*, 95, 1143-1156.

Stern, H.S. and Cressie, N. (2000) Posterior predictive model checks for disease mapping models, *Statistics in Medicine*, 19, 2377-2397.

Smith, A.F.M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte-Carlo methods, *Journal of the Royal Statistical Society*, Ser. B, 55,3-23.

Van der Vaart, A. and Wellner, J. A. (1996) *Weak convergence and empirical processes* Springer, New York.